General linear methods for differential-algebraic equations of index one and two

Philippe Chartier

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

General Linear Methods for Differential-Algebraic Equations of Index One and Two

Philippe CHARTIER

N° 1968
Août 1993

PROGRAMME 1

Architectures parallèles,
bases de données,
réseaux et systèmes distribués

General Linear Methods for Differential-Algebraic Equations of Index one and two

Philippe Chartier*

Programme 6 — Calcul scientifique, modélisation et logiciel numérique
Projet Aladin

Rapport de recherche n° 1968 — Aout 1993 — 24 pages

Abstract: In this contribution, we consider the convergence of the general linear methods of J.C. Butcher when applied to differential algebraic equations of index one and two. For index two systems, the study is limited to stiffly accurate methods. However, preliminary lemmas and theorems are established in a more general context.

Key-words: index-one and index-two differential algebraic equations, general linear methods, stiff accuracy.

(Résumé : tsvp)

*SIMULOG, 1 rue James Joule, 78182 ST QUENTIN YVELINES CEDEX and IRISA, Campus de Beaulieu, 35042 RENNES CEDEX

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : (33) 99 84 71 00 – Télécopie : (33) 99 38 38 32

Méthodes de J.C. Butcher appliquées aux équations différentielles algébriques

Résumé : Nous étudions ici la convergence des méthodes "general linear" de J.C. Butcher appliquées aux équations algébro-différentielles d'indices un et deux. Pour l'indice deux, l'étude n'est poussée jusqu'à son terme que dans le cas des méthodes "stiffly accurate". Les résultats préliminaires sont cependant établis dans un cadre général.

Mots-clé : équations algébro-différentielles d'indice un et deux, méthodes "general linear", méthodes "stiffly accurate".

1 Introduction

In this paper, we are concerned with the convergence of a subclass of Butcher's "general linear methods" (GLM's) when applied to differential algebraic systems of index one and two. GLM's were introduced almost 25 years ago as a unifying tool for theoretical studies (see [2, 8] for an introduction to GLM's) but have never been promoted to practical methods since. However, the methods introduced in [6] are easily formulated as GLM's while no other format enables such an easy presentation. In addition to this, Butcher himself has recently characterized a family of GLM's which has "considerable potential for efficient implementation" (see [3, 4, 5]). These methods, called DIMSIM's of type II and IV share with methods $\mathcal{M}(k, r_k)$ of [6] good stability properties and a high stage-order. Both families are consequently good candidates for the integration of DAE's. In conclusion, it seems especially appropriate to derive convergence results for DAE's within the framework of GLM's. Such results have been already developped for Runge-Kutta's methods [9, 7] and multistep methods [1, 9] and the proofs given here are naturally inspired by the work of [1, 9, 7]. In this paper, we will restrict ourselves to general linear methods whose "correct value function" is easy to interprete and leads to a straighforward generalization of the "direct approach". In Section 2, we propose a definition of stiff accuracy for GLM's and we derive convergence results for stable and strictly stable methods at infinity. Section 3 deals with the index 2 case, that raises naturally more difficult questions. Existence and uniqueness of the solution of the non-linear system to be solved at each step is first shown. Then, the influence of perturbations is studied and rough local error estimates are given. Finally, the convergence of stiffly accurate methods is considered, and it is proved that for such GLM's, the orders of convergence are $\min(p, q+1)$ for the differential component and $\min(p-1, q)$ for the algebraic component, if p denotes the classical order and q the stage-order. In Section 5, we finally show that the methods defined in [6] behave as expected, and first numerical results confirm theoretical investigations.

2 General linear methods for problems of index 1

We first consider systems of the form

$$\begin{cases} y' = f(y,z) \in \mathbb{R}^{m_1} \\ 0 = g(y,z) \in \mathbb{R}^{m_2} \end{cases} \tag{1}$$

with *consistent* initial values, i.e. $g(y_0, z_0) = 0$. We assume that f and g are smooth enough. For this problem to be of index 1, we further assume that

$$g_z(y,z) = \frac{\partial g}{\partial z}(y,z) \text{ is non-singular} \tag{2}$$

and of bounded inverse in a neighborhood of the solution of (1). More precisely, we suppose that there exist two constant K and C, such that

$$\forall (y,z) \in \mathbb{R}^{m_1} \times \mathbb{R}^{m_2}, \max(\|y - y(x)\|, \|z - z(x)\|) \leq K \Longrightarrow \|(\frac{\partial g}{\partial z}(y,z))^{-1}\| \leq C. \tag{3}$$

2.1 Preliminaries

In the sequel, a general linear method will be denoted by the set of matrices $(A, B, \tilde{A}, \tilde{B}, c)$ defined as follows :

$$A = (a_{i,j}) \in \mathbb{R}^{k\times k},\ B = (b_{i,j}) \in \mathbb{R}^{k\times s},\ \tilde{A} = (\tilde{a}_{i,j}) \in \mathbb{R}^{s\times k},\ \tilde{B} = (\tilde{b}_{i,j}) \in \mathbb{R}^{s\times s},\ c = (c_1, \cdots, c_s)^T \quad (4)$$

Let us define the vectors of *internal stages* $v_n = (v_{n,1}^T, \cdots, v_{n,s}^T)^T$ and of *external stages* $u_n = (u_{n,1}^T, \cdots, u_{n,k}^T)^T$. Vector u_n approximates $\mathcal{U}(x_n, h)$ where $\mathcal{U}(x, h)$ is the *correct value function* : $\mathcal{U}(x, h)$ gives the interpretation of the method. With those notations, the *forward step procedure* proceeds as follows :

$$\left\{ \begin{array}{rcl} u_{n+1} & = & (A \otimes I)u_n + h(B \otimes I)F(v_n) \\ v_n & = & (\tilde{A} \otimes I)u_n + h(\tilde{B} \otimes I)F(v_n) \end{array} \right. \quad (5)$$

where F denotes

$$F(v_n) = \left(\begin{array}{c} f(x_n + c_1 h, v_{n,1}) \\ \vdots \\ f(x_n + c_s h, v_{n,s}) \end{array} \right) \quad (6)$$

and I is the $m \times m$ identity matrix if m is the dimension of the system of ordinary differential equations under consideration. In this paper, we will restrict ourselves to the case where the correct value function $\mathcal{U}(x, h)$ is composed only of values of the exact solution evaluated at different points. General linear methods with such exact value function include as particular cases Runge-Kutta methods, multistep methods, multistep collocation methods, one-block methods (see Example 1) and DIMSIM's.

Example 1 *For one-block methods, internal stages and external stages are given by the same equation*

$$Y_{n+1} = (A_{Block} \otimes I)Y_n + h(B_{Block} \otimes I)F(Y_n) \quad (7)$$

where Y_n *approximates* $(y(x_n + (c_1^{Bl} - 1)h)^T, \cdots, y(x_n + (c_k^{Bl} - 1)h)^T)^T$, $c^{Block} \in \mathbb{R}^k$. *We have immediatly* $\tilde{A} = A = A_{Block}$, $\tilde{B} = B = B_{Block}$ *and* $c = c^{Block}$. *The correct value function is simply* $\mathcal{U}(x, h) = (y(x + (c_1 - 1)h)^T, \cdots, y(x + (c_k - 1)h)^T)^T$.

For ordinary differential equations, the matrix $M(z) = A + zB(I - z\tilde{B})^{-1}\tilde{A}$, usually called the *amplification matrix*, determines the stability region. In the case of differential algebraic equations, the limit of this matrix at infinity plays a decisive role. For non-singular $\tilde{B}$, it is given by

$$M(\infty) = A - B\tilde{B}^{-1}\tilde{A} \quad (8)$$

Definition 1 *An implicit general linear method with* $\tilde{B}$ *non-singular is stable at infinity iff*

$$\left\{ \begin{array}{l} \rho(M(\infty)) \leq 1 \\ \forall \lambda \in Sp\{M(\infty)\}, |\lambda| = 1 \Rightarrow \lambda \textit{ is non-defective} \end{array} \right. \quad (9)$$

2.2 The direct approach

A problem of index 1 may be seen as the limit when $\epsilon \to 0$ of the following ordinary differential equation

$$\begin{cases} y' = f(y,z) \\ \epsilon z' = g(y,z). \end{cases} \tag{10}$$

Now, when applying a general linear method to (10), we get :

$$\begin{cases} v_n^y &= (\tilde{A} \otimes I_{m_1}) u_n^y + h(\tilde{B} \otimes I_{m_1}) F(v_n^y, v_n^z) \\ \epsilon v_n^z &= \epsilon(\tilde{A} \otimes I_{m_2}) u_n^z + h(\tilde{B} \otimes I_{m_2}) G(v_n^y, v_n^z) \\ u_{n+1}^y &= (A \otimes I_{m_1}) u_n^y + h(B \otimes I_{m_1}) F(v_n^y, v_n^z) \\ \epsilon u_{n+1}^z &= \epsilon(A \otimes I_{m_2}) u_n^z + h(B \otimes I_{m_2}) G(v_n^y, v_n^z) \end{cases} \tag{11}$$

where

$$F(v_n^y, v_n^z) = \begin{pmatrix} f(v_{n,1}^y, v_{n,1}^z) \\ \vdots \\ f(v_{n,s}^y, v_{n,s}^z) \end{pmatrix} \tag{12}$$

and similarly for $G(v_n^y, v_n^z)$. We now suppose that matrix $\tilde{B}$ is non-singular. From (11), we have

$$hG(v_n^y, v_n^z) = \epsilon[(\tilde{B}^{-1} \otimes I_{m_2}) v_n^z - (\tilde{B}^{-1}\tilde{A} \otimes I_{m_2}) u_n^z] \tag{13}$$

so that

$$\epsilon u_{n+1}^z = \epsilon(A \otimes I_{m_2}) u_n^z + \epsilon(B \otimes I_{m_2})[(\tilde{B}^{-1} \otimes I_{m_2}) v_n^z - (\tilde{B}^{-1}\tilde{A} \otimes I_{m_2}) u_n^z] \tag{14}$$

This last relation defines u_{n+1}^z independently of ϵ. In the sequel, we consequently take $\epsilon = 0$ ("direct approach") and get the scheme for (1) :

$$\begin{cases} v_n^y &= (\tilde{A} \otimes I_{m_1}) u_n^y + h(\tilde{B} \otimes I_{m_1}) F(v_n^y, v_n^z) \\ 0 &= G(v_n^y, v_n^z) \\ u_{n+1}^y &= (A \otimes I_{m_1}) u_n^y + h(B \otimes I_{m_1}) F(v_n^y, v_n^z) \\ u_{n+1}^z &= (M(\infty) \otimes I_{m_2}) u_n^z + (B\tilde{B}^{-1} \otimes I_{m_2}) v_n^z \end{cases} \tag{15}$$

Remark 1 *u_n^y denotes a vector of approximations to the y-component of the solution. u_n^z is defined accordingly. Similarly, $\mathcal{U}^y(x,h)$ stands for the correct-value function of the y-component and $\mathcal{U}^z(x,h)$ for the correct-value function of the z-component.*

Example 2 *The recursion of multistep methods*

$$\sum_{i=0}^{k} \alpha_i y_{n+i} = h \sum_{i=0}^{k} \beta_i f(y_{n+i}) \tag{16}$$

can be easily rewritten as a one-step formula

$$u_{n+1} = \Big(\underbrace{\begin{pmatrix} -\frac{\alpha_{k-1}}{\alpha_k} & \cdots & \cdots & -\frac{\alpha_0}{\alpha_k} \\ 1 & & & 0 \\ & \ddots & & \vdots \\ & & 1 & 0 \end{pmatrix}}_{A \in \mathbb{R}^{k\times k}} \otimes I\Big) u_n + h\Big(\underbrace{\begin{pmatrix} \frac{\beta_k}{\alpha_k} & \cdots & \cdots & \frac{\beta_0}{\alpha_k} \\ 0 & \cdots & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 \end{pmatrix}}_{B \in \mathbb{R}^{k\times(k+1)}} \otimes I\Big) F(v_n) \qquad (17)$$

$$v_n = \Big(\underbrace{\begin{pmatrix} -\frac{\alpha_{k-1}}{\alpha_k} & \cdots & \cdots & -\frac{\alpha_0}{\alpha_k} \\ 1 & & & 0 \\ & \ddots & & \vdots \\ & & 1 & 0 \\ & & & 1 \end{pmatrix}}_{\tilde{A} \in \mathbb{R}^{(k+1)\times k}} \otimes I\Big) u_n + h\Big(\underbrace{\begin{pmatrix} \frac{\beta_k}{\alpha_k} & \cdots & \cdots & \frac{\beta_0}{\alpha_k} \\ 0 & \cdots & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & 0 \end{pmatrix}}_{\tilde{B} \in \mathbb{R}^{(k+1)\times(k+1)}} \otimes I\Big) F(v_n) \qquad (18)$$

where $u_n = (y_{n+k-1}^T, \cdots, y_n^T)^T$ *and* $v_n = (y_{n+k}^T, \cdots, y_n^T)^T$. *Clearly,* $c = (k, k-1, \cdots, 0)^T$. *It should be emphasized that the direct approach for the corresponding general linear method leads to* $g(y_n, z_n) = 0$ *for all* n *and is consequently equivalent to the* **indirect approach** *for the multistep method!*

2.3 Stiff accuracy for general linear methods

Stiff accuracy is determinant for Runge-Kutta's methods with respect to the order of convergence. In the following definition, we extend this notion to general linear methods.

Definition 2 *A general linear method* $(A, B, \tilde{A}, \tilde{B}, c)$ *whose correct value function is composed only of values of the exact solution is called stiffly accurate if for all* $l \in \{1, \cdots, k\}$ *there exists* $m \in \{1, \cdots, s\}$ *such that*

$$\forall j \in \{1, \cdots, k\},\ a_{l,j} = \tilde{a}_{m,j} \qquad (19)$$

$$\forall j \in \{1, \cdots, s\},\ b_{l,j} = \tilde{b}_{m,j}. \qquad (20)$$

For a stiffly accurate general linear method, one has

$$\begin{aligned} (B\tilde{B}^{-1})_{l,j} &= \sum_{r=1}^{s} b_{l,r}\tilde{b}^{-1}_{r,j} \qquad (21) \\ &= \sum_{r=1}^{s} \tilde{b}_{m,r}\tilde{b}^{-1}_{r,j} \\ &= \delta_{m,j} \end{aligned}$$

where δ is the Kronecker symbol, and

$$\begin{aligned} (A - B\tilde{B}^{-1}\tilde{A})_{l,j} &= a_{l,j} - \sum_{r=1}^{s} (B\tilde{B}^{-1})_{l,r}\tilde{a}_{r,j} \qquad (22) \\ &= a_{l,j} - \sum_{r=1}^{s} \delta_{m,r}\tilde{a}_{r,j} \\ &= a_{l,j} - \tilde{a}_{m,j} = 0 \end{aligned}$$

Hence, the l^{th} row of the fourth equation of (15) reduces to

$$\begin{aligned} u^z_{n+1,l} &= [(0,\cdots,0)\otimes I_{m_2}]u^z_n + [(\delta_{m,1},\cdots,\delta_{m,s})\otimes I_{m_2}]v^z_n \\ &= v^z_{n,m} \end{aligned} \tag{23}$$

The same relation obviously holds for $u^y_{n+1,l}$. In particular, we have

$$\forall l \in \{1,\cdots,k\},\ \ g(u^y_{n+1,l}, u^z_{n+1,l}) = 0, \tag{24}$$

so that the approximation provided by the method lies on the manifold $g(y,z)=0$. This means that the numerical solution of (1) by a stiffly accurate GLM coincides with those of the same GLM applied to the reformulated problem

$$\begin{cases} y' = f(y, \mathcal{G}(y)) \\ z = \mathcal{G}(y) \end{cases} \tag{25}$$

which is obtained by application of the Implicit Function Theorem to (1).

Example 3 *A Runge Kutta method is characterized by its Butcher's array*

$$\begin{array}{c|c} c_{RK} & A_{RK} \\ \hline & b^T \end{array}$$

where A_{RK} is $s \times s$ real matrix or equivalently by the equations

$$Y_{n,i} = y_n + h\sum_{j=1}^{s} a^{RK}_{i,j} f(Y_{n,j}) \tag{26}$$

$$y_{n+1} = y_n + h\sum_{j=1}^{s} b_j f(Y_{n,j}) \tag{27}$$

With $u_n = y_n$, $v_n = (Y^T_{n,1},\cdots,Y^T_{n,s})^T$, we have immediately

$$A = 1,\ \tilde{A} = \mathbb{1} \in \mathbb{R}^s,\ \tilde{B} = A_{RK},\ B = b^T,\ c = c_{RK} \tag{28}$$

and $\mathcal{U}(x,h) = y(x)$. Since a stiffly accurate Runge-Kutta method satisfies

$$\forall j \in \{1,\cdots,s\},\ a^{RK}_{s,j} = b_j, \tag{29}$$

its formulation as a general linear method is obviously stiffly accurate (with respect to Definition 2). Conversely, if a stiffly accurate general linear method can be written as a Runge-Kutta method, then it satisfies (29) (nothing prevents us to exchange two rows of A_{RK} and the corresponding components of c_{RK}).

2.4 Convergence

For the direct approach, we have the following convergence result, which is a straightforward generalization of Theorem 1.1 pp. 408 of [9].

Theorem 1 *Suppose that (1) satisfies (2) and that the initial conditions are consistent. Consider an implicit general linear method $(A, \tilde{A}, B, \tilde{B}, c)$ of order p, stage order q (with $p \geq q$), stable at the origin and having a non-singular matrix $\tilde{B}$, and suppose that the error of the starting procedure $\Phi(h)$ satisfies*

$$\Phi^y(h) - \mathcal{U}^y(x_0, h) = \mathcal{O}(h^p) \tag{30}$$
$$\Phi^z(h) - \mathcal{U}^z(x_0, h) = \mathcal{O}(h^{q+1}) \tag{31}$$

Then the global error of the integration procedure (15) satisfies

$$u_n^y - \mathcal{U}^y(x_n, h) = \mathcal{O}(h^p) \tag{32}$$
$$u_n^z - \mathcal{U}^z(x_n, h) = \mathcal{O}(h^r) \tag{33}$$

whenever $nh \leq Const$, where
a) if the method is stiffly accurate, then $u_n^z - \mathcal{U}^z(x_n, h) = \mathcal{O}(h^p)$.
b) $r = \min(p, q+1)$ if the method is stable at infinity and $1 \notin Sp\{M(\infty)\}$.
c) $r = \min(p-1, q)$ if the method is stable at infinity and $1 \in Sp\{M(\infty)\}$.
d) If the method is not stable at infinity, the solution u_n^z diverges.

proof : Part a) has already been discussed. For the remaining cases, we introduce the vectors

$$\mathcal{V}^y(x, h) = (y(x + c_1 h)^T, \cdots, y(x + c_s h)^T)^T \tag{34}$$

and $\mathcal{V}^z(x, h)$ defined accordingly for the z component. We have

$$\mathcal{V}^z(x_n, h) = (\tilde{A} \otimes I_{m_2})\mathcal{U}^z(x_n, h) + h(\tilde{B} \otimes I_{m_2})\mathcal{V}'^z(x_n, h) + \mathcal{O}(h^{q+1}) \tag{35}$$
$$\mathcal{U}^z(x_{n+1}, h) = (A \otimes I_{m_2})\mathcal{U}^z(x_n, h) + h(B \otimes I_{m_2})\mathcal{V}'^z(x_n, h) + \mathcal{O}(h^{q+1}) \tag{36}$$

Since $\tilde{B}$ is non-singular, we can compute $\mathcal{V}'^z(x_n, h)$ from (35) and insert it into (36). This gives

$$\mathcal{U}^z(x_{n+1}, h) = (M(\infty) \otimes I_{m_2})\mathcal{U}^z(x_n, h) + (B\tilde{B}^{-1} \otimes I_{m_2})\mathcal{V}^z(x_n, h) + \mathcal{O}(h^{q+1}) \tag{37}$$

We then denote the global error for the z component by $\Delta U_n^z = u_n^z - \mathcal{U}^z(x_n, h)$ and we denote $\Delta V_n^z = v_n^z - \mathcal{V}^z(x_n, h)$. Substracting (37) from the fourth equation of (15) yields

$$\Delta U_{n+1}^z = (M(\infty) \otimes I_{m_2})\Delta U_n^z + (B\tilde{B}^{-1} \otimes I_{m_2})\Delta V_n^z + \mathcal{O}(h^{q+1}) \tag{38}$$

Now, since the first three equations of (15) define u_{n+1}^y independently of u_n^z, u_{n+1}^y coincides with the solution of (25) by the same method. In particular, we have

$$\Delta U_n^y = e_p(x)h^p + \mathcal{O}(h^{p+1}) \tag{39}$$

due to convergence results of GLM's for ODE's and the existence of an asymptotic expansion. Now, writting equation (35) for the y-component, and substracting the resulting formula from (15), we get

$$\Delta V_n^y = (\tilde{A} \otimes I_{m_1})\Delta U_n^y + h(\tilde{B} \otimes I_{m_1})\{F(v_n^y, \mathcal{G}(v_n^y)) - F(\mathcal{V}^y(x_n, h), \mathcal{G}(\mathcal{V}^y(x_n, h))\} + \mathcal{O}(h^{q+1}) \quad (40)$$

Using a Lipschitz condition for F and the uniform boundness of $(g_z)^{-1}$ leads to $\Delta V_n^y = \mathcal{O}(h^\nu)$ with $\nu = \min(p, q+1)$. Now, the second equation of (15) gives $\Delta V_n^z = \mathcal{O}(h^\nu)$ and (38) becomes

$$\Delta U_{n+1}^z = (M(\infty) \otimes I_{m_2})\Delta U_n^z + \mathcal{O}(h^\nu). \quad (41)$$

Writting the recursion in full leads to

$$\Delta U_n^z = (M(\infty)^n \otimes I_{m_2}) \underbrace{(\Phi^z(h) - \mathcal{U}^z(x_0, h))}_{\delta_0 = \mathcal{O}(h^{q+1})} + \sum_{i=1}^{n} (M(\infty)^{n-i} \otimes I_{m_2})\delta_i \text{ with } \delta_i = \mathcal{O}(h^\nu) \quad (42)$$

This proves the statement when 1 belongs to the spectrum of $M(\infty)$. If 1 does not belong to the spectrum of $M(\infty)$, we can write (42) by using the Partial Summation of Abel as

$$\begin{aligned} \sum_{i=0}^{n} (M(\infty)^{n-i} \otimes I_{m_2})\delta_i &= [(I - M(\infty)^{n+1})(I - M(\infty))^{-1} \otimes I_{m_2}]\delta_0 \\ &+ \sum_{i=1}^{n} [(I - M(\infty)^{n-i+1})(I - M(\infty))^{-1} \otimes I_{m_2}](\delta_i - \delta_{i-1}) \end{aligned} \quad (43)$$

and the result follows easily by noticing that $(\delta_i - \delta_{i-1}) = \mathcal{O}(h^{\nu+1})$. □

3 General linear methods for problems of index 2

We now consider problems of the form

$$\begin{cases} y' = f(y, z) \in \mathbf{R}^{m_1} \\ 0 = g(y) \in \mathbf{R}^{m_2} \end{cases} \quad (44)$$

with *consistent initial value*, i.e. $g(y_0) = 0$ and $g_y(y_0)f(y_0, z_0) = 0$. We will assume that f and g are smooth enough and that $g_y(y)f_z(y, z)$ is non-singular in a neighborhood of the solution of (44). More precisely, we suppose that there exist two constant K and C, such that

$$\forall (y, z) \in \mathbb{R}^{m_1} \times \mathbf{R}^{m_2}, \max(\|y - y(x)\|, \|z - z(x)\|) \leq K \Longrightarrow \|(\frac{\partial g}{\partial y}\frac{\partial f}{\partial z}(y, z))^{-1}\| \leq C. \quad (45)$$

3.1 The direct approach

Similarly to index 1 problem, a general linear method applied to (44) reads

$$\begin{cases} v_n^y &= (\tilde{A} \otimes I_{m_1})u_n^y + h(\tilde{B} \otimes I_{m_1})F(v_n^y, v_n^z) \\ 0 &= G(v_n^y) \\ u_{n+1}^y &= (A \otimes I_{m_1})u_n^y + h(B \otimes I_{m_1})F(v_n^y, v_n^z) \\ u_{n+1}^z &= (M(\infty) \otimes I_{m_2})u_n^z + (B\tilde{B}^{-1} \otimes I_{m_2})v_n^z \end{cases} \quad (46)$$

3.2 Existence and uniqueness of the numerical solution

We first have to show that the first two equations of (46) have a unique solution. Once this solution is found, we can compute u^y_{n+1} and u^z_{n+1} from the last two equations of (46). Here and in Section 3.3, we will denote $v^y_{n,i}$ and $v^z_{n,i}$ by v^y_i and v^z_i respectively for sake of clarity.

Theorem 2 *Suppose that for a solution $y(x), z(x)$ of (44) the following estimates hold :*

$$u^y_n - \mathcal{U}^y(x_n, h) = \mathcal{O}(h),\ G(u^y_n) = \mathcal{O}(h^2) \tag{47}$$

If $\tilde{B}$ is non-singular and if $(g_y f_z)(y, z)$ is non-singular in a neighborhood of the solution, then the nonlinear system

$$\left.\begin{array}{l} v^y_i = \sum_{j=1}^k \tilde{a}_{i,j} u^y_{n,j} + h \sum_{j=1}^s \tilde{b}_{i,j} f(v^y_j, v^z_j) \\ 0 = g(v^y_i) \end{array}\right\} i = 1, \cdots, s \tag{48}$$

has a locally unique solution for h sufficiently small, satisfying $v^y_i - y(x_n) = \mathcal{O}(h)$ and $v^z_i - z(x_n) = \mathcal{O}(h)$.

proof : We put

$$\eta_i = \sum_{i=1}^k \tilde{a}_{i,j} u^y_{n,j} \tag{49}$$

and we define ζ_i close to $z(x_n)$ such that $g_y(\eta_i) f(\eta_i, \zeta_i) = 0$. Let us first notice that $\sum_{i=1}^k \tilde{a}_{i,j} = 1$ due to the preconsistency conditions, so that

$$\begin{aligned} \eta_i - y(x_n) &= \sum_{j=1}^k \tilde{a}_{i,j}(u^y_{n,j} - y(x_n)) \\ &= \sum_{j=1}^k \tilde{a}_{i,j}(u^y_{n,j} - \mathcal{U}^y_j(x_n, h)) + \mathcal{O}(h) \\ &= \mathcal{O}(h) \end{aligned} \tag{50}$$

It comes further,

$$\begin{aligned} g(\eta_i) &= g(y(x_n)) + g_y(y(x_n))(\eta_i - y(x_n)) + \mathcal{O}(\|\eta_i - y(x_n)\|^2) \\ &= g_y(y(x_n)) \sum_{j=1}^k \tilde{a}_{i,j}(u^y_{n,j} - y(x_n)) + \mathcal{O}(h^2) \\ &= \sum_{j=1}^k \tilde{a}_{i,j} g_y(y(x_n))(u^y_{n,j} - y(x_n)) + \mathcal{O}(h^2) \\ &= \sum_{j=1}^k \tilde{a}_{i,j}\{g(u^y_{n,j}) - g(y(x_n)) + \mathcal{O}(\|u^y_{n,i} - y(x_n)\|^2)\} + \mathcal{O}(h^2) \end{aligned} \tag{51}$$

From (47) we consequently get

$$g(\eta_i) = \mathcal{O}(h^2) \tag{52}$$

Now, we write (48) as

$$v_i^y - \eta_i - h\sum_{j=1}^{s} \tilde{b}_{i,j} f(v_j^y, v_j^z) = 0 \tag{53}$$

$$\int_0^1 g_y(\eta_i + \tau(v_i^y - \eta_i))d\tau . \sum_{j=1}^{s} \tilde{b}_{i,j} f(v_j^y, v_j^z) + \frac{1}{h} g(\eta_i) = 0 \tag{54}$$

For $h = 0$, the values

$$v_i^y = \eta_i = \eta_c = y(x_n) \tag{55}$$

$$v_i^z = \zeta_i = \zeta_c = z(x_n) \tag{56}$$

satisfy (53) and (54), since $g(\eta_i(h)) = \mathcal{O}(h^2)$ and $g_y(\eta_i(0))f(\eta_i(0), \zeta_i(0)) = g_y(\eta_c)f(\eta_c, \zeta_c) = 0$. In addition to this, the derivative of (53,54) for $h = 0$ with respect to $(v_{n,1}^y, \cdots, v_{n,s}^y, v_{n,1}^z, \cdots, v_{n,s}^z)$ at point $(\eta_c, \cdots, \eta_c, \zeta_c, \cdots, \zeta_c)$ is of the form

$$\begin{pmatrix} I_{m_1} & \cdots & 0 & \\ \vdots & \ddots & \vdots & 0 \\ 0 & \cdots & I_{m_1} & \\ & \mathrm{L} & & (\tilde{\mathrm{B}} \otimes \mathrm{I})\{\mathrm{g_y f_z}(\eta_c, \zeta_c)\} \end{pmatrix}. \tag{57}$$

Since the blocks on the diagonal are non-singular, matrix (57) is non-singular. In fact, using the uniform bound for the inverse of $g_y f_z$ in a neighborhood of the solution, we see that if h is sufficiently small the inverse of matrix (57) is bounded in a neighborhood of the solution, and the result follows from the Implicit Function Theorem. □

3.3 Influence of Perturbations

The following theorem will be the main tool to establish convergence results.

Theorem 3 *Let v_i^y, v_i^z be given by (48) and consider perturbed values $\hat{v}_i^y$, $\hat{v}_i^z$ satisfying*

$$\left.\begin{array}{l} \hat{v}_i^y = \sum_{j=1}^{k} \tilde{a}_{i,j} \hat{u}_{n,j}^y + h\sum_{j=1}^{s} \tilde{b}_{i,j} f(\hat{v}_j^y, \hat{v}_j^z) + h\delta_i \\ 0 = g(\hat{v}_i^y) + \theta_i \end{array}\right\} \; i = 1, \cdots, s. \tag{58}$$

In addition to the assumptions of Theorem 2 suppose that

$$\hat{u}_n^y - u_n^y = \mathcal{O}(h^2),\ \delta = (\delta_1^T, \cdots, \delta_s^T)^T = \mathcal{O}(h),\ \theta = (\theta_1^T, \cdots, \theta_s^T)^T = \mathcal{O}(h^2). \tag{59}$$

Then there exists a h_0 such that for all $h \le h_0$ we have the following estimates

$$\begin{array}{l} \|\Delta V^y\| \le C\,(\|\Delta U_n^y\| + h\|\delta\| + \|\theta\|) \\ \|\Delta V^z\| \le \frac{C}{h}\left(\sum_{j=1}^{k} \|g_y(\hat{u}_{n,k}^y)(\hat{u}_{n,j}^y - u_{n,j}^y)\| + h\|\Delta U_n^y\| + h\|\delta\| + \|\theta\|\right) \end{array} \tag{60}$$

where

$$\Delta V^y = \left((\hat{v}_1^y - v_1^y)^T, \cdots, (\hat{v}_s^y - v_s^y)^T\right)^T, \quad \Delta U_n^y = \left((\hat{u}_{n,1}^y - u_{n,1}^y)^T, \cdots, (\hat{u}_{n,k}^y - u_{n,k}^y)^T\right)^T,$$
$$\|\Delta V^y\| = \max_{1\le j\le s} \|\hat{v}_j^y - v_j^y\|, \qquad \|\Delta U_n^y\| = \max_{1\le j\le k} \|\hat{u}_{n,j}^y - u_{n,j}^y\|,$$

and likewise for the z-component.

The result now follows by noticing that $\Delta\eta = \mathcal{O}(h^2)$ and $u^y_{n,k} - \eta_i = \mathcal{O}(h)$, so that

$$\begin{aligned} \|g(\hat{\eta}_i) - g(\eta_i)\| &\leq \|g_y(\eta_i)(\hat{\eta}_i - \eta_i)\| + C_1\|\hat{\eta}_i - \eta_i)\|^2) \\ &\leq \|g_y(\eta_i)(\hat{\eta}_i - \eta_i)\| + (C_1C_2h)h\|\hat{\eta}_i - \eta_i)\| \\ &\leq \|g_y(\eta_i)(\hat{\eta}_i - \eta_i)\| + hC\|\hat{\eta}_i - \eta_i)\| \end{aligned} \tag{77}$$

where C is independent of the constant involved in the $\mathcal{O}$-term of $\Delta\eta = \mathcal{O}(h^2)$, i.e finally,

$$\begin{aligned} \|g(\hat{\eta}_i) - g(\eta_i)\| &\leq \|g_y(u^y_{n,k})(\hat{\eta}_i - \eta_i)\| + hC\|\hat{\eta}_i - \eta_i)\| \\ &\leq \|\sum_{j=1}^{k} \tilde{a}_{i,j} g_y(u^y_{n,k})(\hat{u}^y_{n,j} - u^y_{n,j})\| + hC_1\|\hat{\eta}_i - \eta_i)\|. \end{aligned}$$

□

Remark 2 *It must be emphasized that the constant C in (60) is independent of the constants involved in $\mathcal{O}$-terms of the assumptions of Theorem 3. For a sufficiently small h_0, this constant depend only on bounds of the derivatives of f and g, uniformly for $h \leq h_0$.*

3.4 Projections P and Q

In order to study the local error, we focus on the following projections that constitute an important tool of convergence proofs.

Definition 3 *For given y, z for which $(g_y f_z)(y,z)$ is non-singular we define the projections*

$$Q = (f_z(g_y f_z)^{-1} g_y)(y,z), \tag{78}$$

$$P = I - Q. \tag{79}$$

Interpretation : let us consider the following example

$$\begin{aligned} f: \quad \mathbf{R}^3 &\longrightarrow \mathbf{R}^2 \\ (y_1, y_2, z) &\longmapsto f(y_1, y_2, z) = (f_1(y_1, y_2, z), f_2(y_1, y_2, z))^T \end{aligned} \tag{80}$$

$$\begin{aligned} g: \quad \mathbf{R}^2 &\longrightarrow \mathbf{R} \\ (y_1, y_2) &\longmapsto g(y_1, y_2) = g(y_1, y_2) \end{aligned} \tag{81}$$

and let us assume that $g_y f_z$ is non-singular at point (y_1^0, y_2^0, z^0). The equation $g(y) = 0$ defines a curve $\mathcal{G}$ in the affine space (O, y_1, y_2) (see Figure 1). The tangent space $T_{\mathcal{G}}$ to $\mathcal{G}$ at point y^0 is the kernel of $g_y(y^0)$,

$$T_{\mathcal{G}} = Ker(g_y(y^0)). \tag{82}$$

In fact, it is here the set of vectors $w = (w_1, w_2)^T$ such that

$$\left.\frac{\partial g}{\partial y_1}\right|_{y=y^0} w_1 + \left.\frac{\partial g}{\partial y_2}\right|_{y=y^0} w_2 = 0 \tag{83}$$

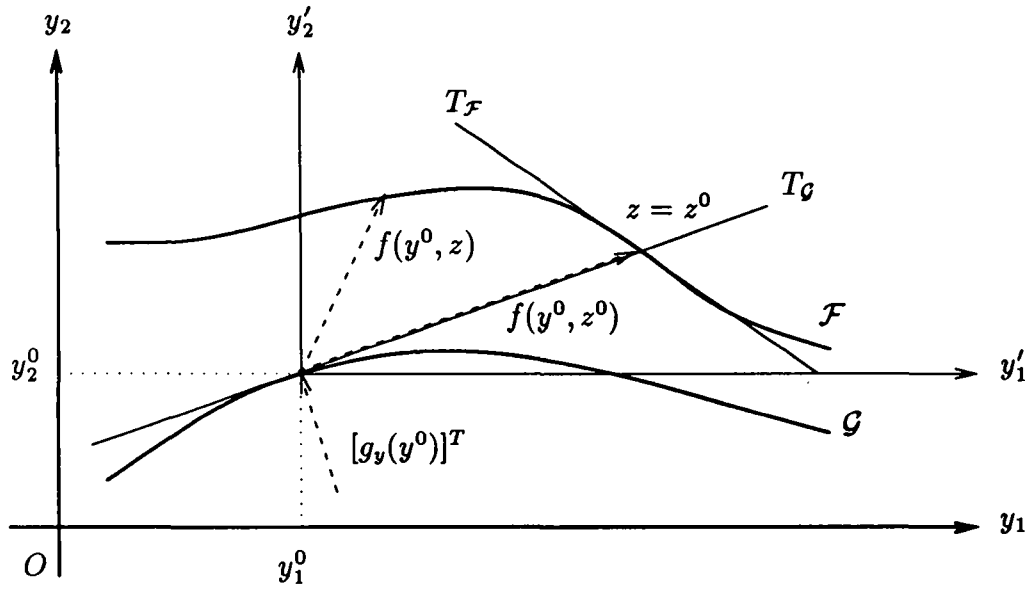

Figure 1: Geometric interpretation of $P(x)$ and $Q(x)$

and it is represented on Figure 1 by a line orthogonal to $[g_y(y^0)]^T$ at point y^0. Similarly, the equation $y' = f(y^0, z)$ defines a parametric curve $\mathcal{F}$, and the tangent space $T_{\mathcal{F}}$ to $\mathcal{F}$ at point z^0 is the image of $f_z(y^0, z^0)$,

$$T_{\mathcal{F}} = Im(f_z(y^0, z^0)). \tag{84}$$

Here, z^0 is taken to satisfy the "hidden constraint" $g_y(y^0)f(y^0, z^0) = 0$. With f defined as above, $T_{\mathcal{F}}$ is made of vectors colinear to $f_z(y^0, z^0)$, and it is represented by the tangent line to $\mathcal{F}$ at point $z = z^0$ (see Figure 1). Considering Figure 1, we see that the space $T_{\mathcal{F}}$ can be interpreted as the set of directions in which z bring the solution to the constraint $g(y) = 0$. Furthermore, P is then to be seen as the projection onto $T_{\mathcal{G}}$ along $T_{\mathcal{F}}$ and Q as the projection onto $T_{\mathcal{F}}$ along $T_{\mathcal{G}}$.

3.5 Local error

Let us now consider two initial vectors $u_0^y = \mathcal{U}^y(x, h)$ and $u_0^z = \mathcal{U}^z(x, h)$ on the exact solution and let us denote u_1^y and u_1^z the numerical solution of the GLM (46) after one step. We define the local error by

$$\delta u_h^y = u_1^y - \mathcal{U}^y(x + h, h), \tag{85}$$

$$\delta u_h^z = u_1^z - \mathcal{U}^z(x + h, h). \tag{86}$$

Theorem 4 *Suppose that a zero-stable GLM $(A, B, \tilde{A}, \tilde{B}, c)$ of order of consistency p and with non-singular matrix $\tilde{B}$ satisfies*

$$\mathcal{U}(x + h, h) - (A \otimes I)\mathcal{U}(x, h) - h(B \otimes I)\mathcal{V}'(x, h) = \mathcal{O}(h^{q+1}) \tag{87}$$

$$\mathcal{V}(x, h) - (\tilde{A} \otimes I)\mathcal{U}(x, h) - h(\tilde{B} \otimes I)\mathcal{V}'(x, h) = \mathcal{O}(h^{q+1}) \tag{88}$$

with $p \geq q \geq 1$. Then we have

$$\delta u_h^y = \mathcal{O}(h^{\min(p,q+1)}), \quad (E \otimes I_{m_1})\delta u_h^y = \mathcal{O}(h^{\min(p+1,q+1)}), \tag{89}$$

$$\{P(x)\}\delta u_h^y = \mathcal{O}(h^{\min(p,q+2)}), \quad \{P(x)\}(E \otimes I_{m_1})\delta u_h^y = \mathcal{O}(h^{\min(p+1,q+2)}) \tag{90}$$

and

$$\delta u_h^z = \mathcal{O}(h^q). \tag{91}$$

If in addition, the method is stiffly accurate, then

$$\delta u_h^y = \mathcal{O}(h^{\min(p,q+2)}), \quad (E \otimes I_{m_1})\delta u_h^y = \mathcal{O}(h^{\min(p+1,q+2)}). \tag{92}$$

proof : Let us consider the following auxiliary system

$$u_{aux}^y = (A \otimes I_{m_1})\mathcal{U}^y(x,h) + h(B \otimes I_{m_1})F(v_{aux}^y, \mathcal{V}^z(x,h)) \tag{93}$$

$$v_{aux}^y = (\tilde{A} \otimes I_{m_1})\mathcal{U}^y(x,h) + h(\tilde{B} \otimes I_{m_1})F(v_{aux}^y, \mathcal{V}^z(x,h)). \tag{94}$$

By definition of the order of consistency of a GLM applied to an ordinary differential equation, we have

$$u_{aux}^y - \mathcal{U}^y(x+h,h) = d_p(x)h^p + \mathcal{O}(h^{p+1}) \tag{95}$$

where $d_p(x)$ depends on $f(y(x), z(x))$. We further have

$$u_1^y - u_{aux}^y = h(B \otimes I_{m_1}) \underbrace{(F(v^y, v^z) - F(v_{aux}^y, \mathcal{V}^z(x,h)))}_{D} \tag{96}$$

D can be divided into three parts :

$$D = \underbrace{F(v^y, v^z) - F(\mathcal{V}^y(x,h), v^z)}_{D_1} + \underbrace{F(\mathcal{V}^y(x,h), v^z) - F(\mathcal{V}^y(x,h), \mathcal{V}^z(x,h))}_{D_2} \tag{97}$$

$$+ \underbrace{F(\mathcal{V}^y(x,h), \mathcal{V}^z(x,h)) - F(v_{aux}^y, \mathcal{V}^z(x,h))}_{D_3}. \tag{98}$$

A Lipschitz condition for f and application of Theorem 3 with $\hat{v}^y = \mathcal{V}^y(x,h)$, $\delta = \mathcal{O}(h^q)$ and $\theta = 0$ give

$$D_1 = \mathcal{O}(h^{q+1}). \tag{99}$$

The same Lipschitz condition for f and relation (88) with $\mathcal{V}^y$ give

$$D_3 = \mathcal{O}(h^{q+1}). \tag{100}$$

As for D_2 we have

$$D_2 = \{f_z(\mathcal{V}_i^y(x,h), \mathcal{V}_i^z(x,h))\}_{i=1,\dots,s}(v^z - \mathcal{V}^z(x,h)) + \mathcal{O}(\|v^z - \mathcal{V}^z(x,h)\|^2). \tag{101}$$

Theorem 3 gives $\|v^z - \mathcal{V}^z(x,h)\| = \mathcal{O}(h^q)$. From $\mathcal{V}_i^y(x,h) - y(x) = \mathcal{O}(h)$ and $\mathcal{V}_i^z(x,h) - z(x) = \mathcal{O}(h)$, we obtain

$$D_2 = \{f_z(y(x),z(x))\}_{i=1,\cdots,s}(v^z - \mathcal{V}^z(x,h)) + \mathcal{O}(h^{q+1}). \tag{102}$$

Gathering all partial results, we get

$$\delta u_h^y = h\{f_z(y(x),z(x))\}_{i=1,\cdots,s}(B \otimes I_{m_2})(v^z - \mathcal{V}^z(x,h)) + d_p(x)h^p + \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{q+2}) \tag{103}$$

and the result follows from $(E \otimes I_{m_1})d_p(x) = 0$ and $P(x)f_z(y(x),z(x)) = 0$. As for the z-component, it comes immediately

$$\begin{aligned} \delta u_h^z &= (B\tilde{B}^{-1} \otimes I_{m_2})(v^z - \mathcal{V}^z(x,h)) + \mathcal{O}(h^{q+1}) \\ &= \mathcal{O}(h^q). \end{aligned} \tag{104}$$

Now, for stiffly accurate methods, we have

$$G(u_1^y) = 0, \tag{105}$$

so that

$$\forall i \in \{1,\cdots,k\},\quad 0 = g(u_{1,i}^y) - g(\mathcal{U}_i^y(x+h,h)) \tag{106}$$

$$= g_y(\mathcal{U}_i^y(x+h,h))((\delta u_h^y)_i) + \mathcal{O}(\|(\delta u_h^y)_i\|^2). \tag{107}$$

From $g_y(\mathcal{U}_i^y(x+h,h)) = g_y(y(x)) + \mathcal{O}(h)$, we get

$$0 = \{g_y(y(x))\}\delta u_h^y + \mathcal{O}(h\|\delta u_h^y\|). \tag{108}$$

Pre-multiplying by $\{f_z(g_y f_z)^{-1}(y(x),z(x))\}$ leads to

$$\{Q(x)\}\delta u_h^y = \mathcal{O}(h^{\min(p+1,q+2)}) \tag{109}$$

and (92) is obtained from (90) and relation $P(x) + Q(x) = I$. □

3.6 Convergence for the y-component

Due to the form of the scheme, the convergence of the y-component can be treated independently. As already mentioned in the introduction, we only deal with the case of stiffly accurate methods.

Theorem 5 *Suppose that $g_y f_z(y,z)$ is non-singular in a neighborhood of the solution $(y(x),z(x))$ of (44) and that the initial values (y_0,z_0) are consistent. Consider a zero-stable stiffly accurate GLM of order p and stage order q, with $p \geq 2$ and $q \geq 1$. Then the method (46) is convergent of order $r = \min(p,q+1)$, i.e.*

$$u_n^y - \mathcal{U}^y(x_n,h) = \mathcal{O}(h^r) \tag{110}$$

for $x_n - x_0 = nh \leq Const.$

proof : We still denote the global error by $\Delta U_n^y = u_n^y - \mathcal{U}_n^y$. In addition to the numerical solution u_n^y given by the scheme (46), we consider the following auxiliary system at point x_n :

$$\left\{\begin{array}{rcl} \hat{u}_{n+1}^y &=& (A\otimes I_{m_1})\mathcal{U}^y(x_n,h) + h(B\otimes I_{m_1})F(\hat{v}_n^y,\hat{v}_n^z) \\ \hat{v}_n^y &=& (\tilde{A}\otimes I_{m_1})\mathcal{U}^y(x_n,h) + h(\tilde{B}\otimes I_{m_1})F(\hat{v}_n^y,\hat{v}_n^z) \\ 0 &=& G(\hat{v}_n^y) \end{array}\right. \tag{111}$$

and we suppose for the moment that we have

$$\|\Delta U_n^y\| \leq C_0 h^2. \tag{112}$$

From Theorem 4, we have $\delta u_h^y(x_n) = \mathcal{O}(h^r)$. If C_1 denotes the constant involved in the $\mathcal{O}(h^r)$-term, the assumptions of Theorem 3 are satisfied with

$$\|u_{n+1}^y - \hat{u}_{n+1}^y\| = \|\Delta U_{n+1}^y - \delta u_h^y(x_n)\| \leq (C_0 + C_1 h^{r-2})h^2 \leq C_2 h^2. \tag{113}$$

It comes further

$$\begin{array}{rcl} \Delta U_{n+1}^y &=& (u_{n+1}^y - \hat{u}_{n+1}^y) + (\hat{u}_{n+1}^y - \mathcal{U}^y(x_{n+1},h)) \\ &=& (A\otimes I_{m_1})\Delta U_n^y + h\underbrace{(B\otimes I_{m_1})(F(v_n^y,v_n^z) - F(\hat{v}_n^y,\hat{v}_n^z))}_{\Delta F_n} + \delta u_h^y(x_n). \end{array} \tag{114}$$

Using a Lipschitz condition for F, we get

$$\|\Delta F_n\| \leq L\|B\otimes I_{m_1}\|\left(\|(v_n^y - \hat{v}_n^y)\| + \|(v_n^z - \hat{v}_n^z)\|\right). \tag{115}$$

Now, the method being stiffly accurate, we have

$$\forall i \in \{1,\cdots,k\}, g(u_{n,i}^y) - g(\mathcal{U}_i^y(x_n,h)) = 0 \tag{116}$$

so that

$$\begin{array}{rcl} 0 &=& g_y(\mathcal{U}_i^y(x_n,h))(\mathcal{U}_i^y(x_n,h) - u_{n,i}^y) + \mathcal{O}(\|\mathcal{U}_i^y(x_n,h) - u_{n,i}^y\|^2) \\ &=& g_y(u_{n,i}^y)(\mathcal{U}_i^y(x_n,h) - u_{n,i}^y) + \mathcal{O}(h^2\|\mathcal{U}_i^y(x_n,h) - u_{n,i}^y\|) \\ &=& g_y(u_{n,k}^y)(\mathcal{U}_i^y(x_n,h) - u_{n,i}^y) + \mathcal{O}(h\|\mathcal{U}_i^y(x_n,h) - u_{n,i}^y\|). \end{array} \tag{117}$$

Applying Theorem 3 with $\delta = 0$ and $\theta = 0$ and taking into account last equation we obtain the following estimations

$$\left\{\begin{array}{l} \|v_n^y - \hat{v}_n^y\| \leq C_3\|\Delta U_n^y\| \\ \|v_n^z - \hat{v}_n^z\| \leq C_3\|\Delta U_n^y\| \end{array}\right. \tag{118}$$

and finally $\|\Delta F_n\| \leq C_4\|\Delta U_n^y\|$. We have consequently a recursion of the form

$$\Delta U_{n+1}^y = (A\otimes I_{m_1})\Delta U_n^y + h\Delta F_n + \delta u_h^y(x_n) \tag{119}$$

with $\|\Delta F_n\| \leq C_4\|\Delta U_n^y\|$, $\delta u_h^y(x_n) = \mathcal{O}(h^r)$ and $(E\otimes I_{m_1})\delta u_h^y(x_n) = \mathcal{O}(h^{r+1})$, where the last two equations come from Theorem 4. Now, following the same ideas as in Lemma 8.12 and

Theorem 8.13 from [8], pp.394-395, we can assert that there exists a positive constant C_5, independent of C_0 such that

$$\forall n, n.h \leq Const \Longrightarrow |\Delta U_n^y|| \leq C_5 h^r. \tag{120}$$

A careful look at Lemma 8.12 from [8] shows that we have in fact

$$\forall n, n.h \leq Const \Longrightarrow \|\Delta U_{n+1}^y\| \leq C_6 h^r. \tag{121}$$

where C_6 is independent of C_0 (see Remark 2). We can now validate assumption (112) by induction on n, by taking $C_0 = C_6$, and restricting the size of h. Since the assumption (112) is trivially satisfied for $n = 0$, we get from (121), $\|\Delta U_1^y\| \leq C_0 h^r \leq (C_0 h^{r-2}) h^2$ for all $h \leq h_0$, h_0 sufficiently small but independent of n. The same relation holds at point x_n, and the induction is then straighforward. □

4 Convergence for the z-component

The global error for the z-component is easy to estimate. It is essentially given by the local error.

Theorem 6 *Suppose that $g_y f_z(y,z)$ is non-singular in a neighborhood of the solution $(y(x), z(x))$ of (44) and that the initial values (y_0, z_0) are consistent. Consider a zero-stable GLM strictly stable at infinity such that the global error of the y-component is $\mathcal{O}(h^r)$, $g(u_{n,i}) = \mathcal{O}(h^{r+1})$ for all $i = 1, \cdots, k$ and the local error of the z-component is $\mathcal{O}(h^r)$ with $r \geq 2$. Then we have for the global error of the z-component :*

$$u_n^z - \mathcal{U}^z(x_n, h) = \mathcal{O}(h^r) \tag{122}$$

for $x_n - x_0 = nh \leq Const$.

proof : We can write the global error for the z-component as

$$u_{n+1}^z - \mathcal{U}^z(x_{n+1}, h) = (u_{n+1}^z - \hat{u}_{n+1}^z) + (\hat{u}_{n+1}^z - \mathcal{U}^z(x_{n+1}, h)) \tag{123}$$

where $\hat{u}_{n+1}^z$ satisfies the auxiliary system

$$\left\{ \begin{array}{rcl} \hat{u}_{n+1}^z & = & (M(\infty) \otimes I_{m_2}) \mathcal{U}^z(x_n, h) + (B\tilde{B} \otimes I_{m_2}) \hat{v}_n^z \\ \hat{v}_n^y & = & (\tilde{A} \otimes I_{m_1}) \mathcal{U}^y(x_n, h) + h(\tilde{B} \otimes I_{m_1}) F(\hat{v}_n^y, \hat{v}_n^z) \\ 0 & = & G(\hat{v}_n^y) \end{array} \right. . \tag{124}$$

It comes

$$u_{n+1}^z - \mathcal{U}^z(x_{n+1}, h) = (M(\infty) \otimes I_{m_2})(u_n^z - \mathcal{U}^z(x_n, h)) + (B\tilde{B} \otimes I_{m_2})(v_n^z - \hat{v}_n^z) + \delta_h^z(x_n). \tag{125}$$

Taking into account relations $g(u_{n,i}) = \mathcal{O}(h^{r+1})$, $i = 1, \cdots, k$, we can show as in Theorem 5 that

$$\|v_n^z - \hat{v}_n^z\| \leq C\,(\|u_n^y - \mathcal{U}^y(x_n, h)\| + h^r) \tag{126}$$

and the result immediately follows from the hypothesis. □

Remark 3 *Theorem 6 is valid for stiffly or non-stiffly accurate methods as well. For stiffly accurate methods,* $g(u_{n,i}) = \mathcal{O}(h^{r+1})$ *becomes* $g(u_{n,i}) = 0$, *while* $M(\infty) = 0$. *In that case, the proof is even more simple.*

5 Application to $\mathcal{M}(k, r_k)$ methods

In [6] has been introduced a class of parallel one-block methods of orders varying from 2 to 8. If k denotes the size of the block, those methods are constructed in such a way that they are of order k and stage order $k-1$. In addition to this, they satisfy $A = \tilde{A}$ and $B = \tilde{B}$ with non-singular B so that $M(\infty) = 0$. From Theorems 1 and 5, we consequently expect the following orders of convergence for index one and two problems (see Table 1).

Method	Index 1		Index 2	
	y	z	y	z
$\mathcal{M}(k, r_k)$	k	k	k	$k-1$

Table 1: Order of convergence of methods $\mathcal{M}(k, r_k)$ when applied to DAE's

5.1 Numerical verification for index one problems

In order to numerically verify the behaviour of methods $\mathcal{M}(k, r_k)$ on index one problems, we have integrated the following examples. The first one is a slight modification of Kap's problem (see [10])

$$\begin{cases} y' = -(2+\epsilon^{-1})y + \epsilon^{-1}z^2, & y(0) = 1 \\ 0 = y - z(1+z) + e^{-x}, & z(0) = 1 \end{cases}, \tag{127}$$

with exact solution

$$\begin{cases} y(x) = e^{-2x} \\ z(x) = e^{-x} \end{cases}. \tag{128}$$

The second one is a test equation obtained by transformation of a linear optimal control problem with quadratic cost functional

$$\begin{cases} \text{Find y and u in } C^1([0,1],\mathbb{R}^n) \text{ such that,} \\ y' = Ay + Bu + g(x), \quad y(0) = y_0 \\ J(u) = \frac{1}{2}\int_0^1 (y(x)^T C y(x) + u(x)^T D u(x))dx \text{ is minimum} \end{cases}, \tag{129}$$

where C and D are assumed to be positive semi-definite. Such problems are completly solved when $g(x) = 0$, but need to be reformulated as a DAE otherwise (see for example [12], pp.289). It can be shown indeed that y and u are solution of the following two points boundary problem

$$\begin{cases} y' = Ay + Bu + g(x), & y(0) = y_0 \\ v' = -A^T v - Cy, & v(1) = 0 \\ 0 = B^T v + Du \end{cases} \tag{130}$$

which is of index one provided D is definite. If $D = 0$ and $B^T CB$ is positive definite, than (130) is of index 3. We consider here the simple case where A, B, C and D are scalars with the following values

$$a(x) = -1, b(x) = \log(2+x), c(x) = 1 + (1-x)(1-2a-x), d(x) = b(x)^2. \tag{131}$$

and initial condition $y_0 = 1$. The corresponding system (130) has

$$\begin{cases} y(x) = e^{(a-1)x+x^2/2}, \\ u(x) = (x-1)\log(2+x)e^{(a-1)x+x^2/2}, \\ v(x) = (1-x)e^{(a-1)x+x^2/2} \end{cases} \tag{132}$$

as exact solution. In fact, we have integrated here the initial value problem with same exact solution

$$\begin{cases} y' = ay + bu, & y(0) = 1 \\ v' = -av - cu, & v(0) = 1 \\ 0 = bv + du \end{cases} . \tag{133}$$

Method / h		1	0.5	0.25	0.125	0.0625	0.03125	0.015625	0.0078125
$\mathcal{M}(2,r_2)$	y	1.35	1.85	2.24	2.59	2.93	3.28	3.66	4.07
	z	5.70	6.19	6.58	6.93	7.28	7.63	8.00	8.42
$\mathcal{M}(3,r_3)$	y	1.83	2.48	3.11	3.74	4.38	5.06	5.80	6.64
	z	6.17	6.83	7.46	8.08	8.73	9.40	10.14	10.98
$\mathcal{M}(4,r_4)$	y	2.20	3.03	3.92	4.91	6.16	7.18	7.56	8.32
	z	6.54	7.37	8.26	9.26	10.50	11.52	11.90	12.65
$\mathcal{M}(5,r_5)$	y	2.40	3.40	4.50	5.70	7.06	8.74	10.24	12.53
	z	6.75	7.74	8.84	10.05	11.40	13.08	14.59	16.87
$\mathcal{M}(6,r_6)$	y	2.60	3.76	5.08	6.54	8.10	9.66	11.08	12.48
	z	6.95	8.10	9.42	10.88	12.45	14.00	15.42	16.83
$\mathcal{M}(7,r_7)$	y	2.77	4.08	5.60	7.31	9.15	11.05	13.01	15.42
	z	7.11	8.42	9.95	11.65	13.49	15.40	17.35	19.77
$\mathcal{M}(8,r_8)$	y	2.93	4.39	6.14	8.11	10.27	12.53	15.18	17.57
	z	7.28	8.74	10.49	12.46	14.59	16.87	19.52	21.92

Table 2: Number of correct digits (Δ) for Problem (127) with $\epsilon = 10^{-2}$ on $[0, 10]$

5.2 Numerical verification for index two problems

We now consider the case of index two problems and integrate the following two systems. The first one

$$\begin{cases} y_1' = -(2+\epsilon^{-1})y_1 + \epsilon^{-1}y_2^2, & y_1(0) = 1 \\ y_2' = -e^{1-x^2}, & y_2(0) = 1 \\ 0 = y_1 - y_2(1+y_2) + \frac{y_1}{y_2}, & z(0) = 1 \end{cases} \tag{134}$$

is once more a modification of Kap's problem, with exact solution

$$\begin{cases} y_1(x) = e^{-2x} \\ y_2(x) = e^{-x} \\ z(x) = \sqrt{1+x} \end{cases} . \tag{135}$$

The second one is a test example from [11] and has the form

$$\begin{cases} y_1' = -y_1 + sin(\nu x)z + q_1(x) \\ y_2' = -y_2 + cos(\nu x)z + q_2(x) \\ 0 = sin(\nu x)y_1 + cos(\nu x)y_2 + r(x) \end{cases} \tag{136}$$

with $q_1(x) = e^x(2 + \frac{sin(\nu x)}{2-x})$, $q_2(x) = e^x(2 + \frac{cos(\nu x)}{2-x})$, and $r(x) = -e^x(sin(\nu x) + cos(\nu x))$. Its exact solution is simply

$$\begin{cases} y_1(x) = e^x \\ y_2(x) = e^x \\ z(x) = -\frac{e^x}{2-x} \end{cases} . \tag{137}$$

6 Concluding remarks

In this paper, we have established convergence results for general linear methods when applied to DAE's of index one and two. In the case of index one, the proof we give is a direct generalization of the corresponding proof for Runge-Kutta methods. For index two problems, we have studied the existence and uniqueness of the solution of the non-linear system involved at each step, together with the influence of perturbation on this solution, in a general context. Namely, no special assumption has been used for these proofs. These results gave us the opportunity to derive estimations of the local error of GLM's. Stiff accuracy leads here to more stringent estimations. For general case, these estimations are not optimal, and more attention should be paid. For this aim, we intend to use in a future work the series of Butcher.

The main result of this paper is the convergence proof for index two systems. We still do not provide convergence results for non-stiffly accurate GLM's, and this excludes DIMSIM's, but the use of stiff accuracy enables much more simple proofs of convergence results in the case of Runge-Kutta methods, and it is likely that the same holds also for GLM's.

Finally, and this was the original motivation for considering the convergence of GLM's when applied to DAE's, we have applied the theorems derived here to methods $\mathcal{M}(k, r_k)$ of [6]. Since these methods are stiffly accurate, the results obtained are optimal. Numerical tests on index one and two problems show the correctness of our analysis.

Acknowledgements

I am grateful to B. Philippe and M. Crouzeix for their helpful comments.

References

[1] K. BREMAN, S. CAMPBELL, AND L. PETZOLD, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, North-Holland, 1989.

[2] J. BUTCHER, *The numerical analysis of ordinary differential equations. Runge-Kutta and general linear methods.*, Wiley-Interscience, 1987.

[3] J. BUTCHER, *Diagonally Implicit Multi Stage Integration Methods*, App. Numer. Math., (1993). (to appear).

[4] ——, *General linear methods for the parallel solution of ordinary differential equations*, (1993). (in preparation).

[5] J. BUTCHER AND Z. JACKIEWICZ, *Diagonally implicit general linear methods for ordinary differential equations*, (1993). (in preparation).

[6] P. CHARTIER, *L-stable parallel one-block methods for ordinary differential equations.*, SIAM Journal of Numerical Analysis, (1993). (to appear).

[7] E. HAIRER, C. LUBICH, AND M. ROCHE, *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*, vol. 1409 of Lecture Notes in Mathematics, Springer-Verlag, 1989.

[8] E. HAIRER, S. NORSETT, AND G. WANNER, *Solving Ordinary Differential Equations I. Nonstiff Problems*, vol. 1, Springer-Verlag, 1987.

[9] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems.*, vol. 2, Springer-Verlag, 1991.

[10] P. KAPS, *Rosenbrock-type methods*, in Numerical methods for stiff initial value problems., D. G. and J. R., eds., Inst. fur Geometrie und Praktische Mathematik der RWTH Aachen., 1981.

[11] L. PETZOLD, *Numerical solution of differential-algebraic equations*, in Ecoles CEA-EDF-INRIA, Problemes non-lineaires appliques, Systemes algebro-differentiels, jun 1992, pp. 1–19.

[12] E. SONTAG, *Mathematical Control Theory*, Texts in Applied Mathematics, Springer-Verlag, 1990.

Method	Average order		Asymptotic order		Predicted order	
	y	z	y	z	y	z
$\mathcal{M}(2,r_2)$	1.3	1.3	1.4	1.4	2	2
$\mathcal{M}(3,r_3)$	2.3	2.3	2.8	2.8	3	3
$\mathcal{M}(4,r_4)$	2.9	2.9	2.5	2.5	4	4
$\mathcal{M}(5,r_5)$	4.8	4.8	7.6	7.6	5	5
$\mathcal{M}(6,r_6)$	4.7	4.7	4.7	4.7	6	6
$\mathcal{M}(7,r_7)$	6.0	6.0	8.0	8.0	7	7
$\mathcal{M}(8,r_8)$	6.9	6.9	7.9	8.0	8	8

Table 3: Observed orders for Problem (127).

Method / h		0.1	0.05	0.025	0.0125	0.00625	0.003125
$\mathcal{M}(2,r_2)$	y	-0.21	0.37	0.93	1.51	2.09	2.69
	z	-0.21	0.37	0.93	1.51	2.09	2.69
$\mathcal{M}(3,r_3)$	y	0.72	2.03	3.62	3.85	4.60	5.43
	z	0.72	2.03	3.62	3.89	4.62	5.45
$\mathcal{M}(4,r_4)$	y	1.31	2.30	3.34	4.44	5.59	6.76
	z	1.31	2.30	3.34	4.44	5.59	6.76
$\mathcal{M}(5,r_5)$	y	1.52	3.01	4.54	6.08	7.62	9.15
	z	1.52	3.01	4.54	6.08	7.62	9.15
$\mathcal{M}(6,r_6)$	y	2.67	4.22	5.74	7.37	9.08	10.87
	z	2.67	4.22	5.74	7.37	9.08	10.87
$\mathcal{M}(7,r_7)$	y	2.46	4.38	6.38	8.44	10.50	12.61
	z	2.46	4.38	6.38	8.44	10.50	12.61
$\mathcal{M}(8,r_8)$	y	4.06	6.31	8.23	10.30	12.56	14.88
	z	4.06	6.31	8.23	10.30	12.56	14.88

Table 4: Number of correct digits (Δ) for Problem (133) on $[0,5]$.

Method	Average order		Asymptotic order		Predicted order	
	y	z	y	z	y	z
$\mathcal{M}(2,r_2)$	1.9	1.9	2.0	2.0	2	2
$\mathcal{M}(3,r_3)$	3.1	3.1	2.8	2.8	3	3
$\mathcal{M}(4,r_4)$	3.6	3.6	3.9	3.9	4	4
$\mathcal{M}(5,r_5)$	5.1	5.1	5.1	5.1	5	5
$\mathcal{M}(6,r_6)$	5.4	5.4	5.9	5.9	6	6
$\mathcal{M}(7,r_7)$	6.7	6.7	7.0	7.0	7	7
$\mathcal{M}(8,r_8)$	7.2	7.2	7.7	7.7	8	8

Table 5: Observed orders for Problem (133).

Method / h		0.1	0.05	0.025	0.0125	0.00625	0.003125	1.562510^{-3}	7.812510^{-4}
$\mathcal{M}(2,r_2)$	y	0.47	1.17	1.81	2.43	3.04	3.65	4.25	4.86
	z	1.57	2.21	2.68	3.09	3.44	3.77	4.09	4.40
$\mathcal{M}(3,r_3)$	y	1.20	2.10	2.99	3.88	4.78	5.68	6.59	7.49
	z	2.50	3.34	4.13	4.88	5.58	6.24	6.88	7.50
$\mathcal{M}(4,r_4)$	y	1.66	2.91	4.14	5.35	6.56	7.77	8.98	10.18
	z	3.00	4.27	5.57	6.97	9.26	9.03	9.77	10.61
$\mathcal{M}(5,r_5)$	y	2.45	3.95	5.45	6.95	8.46	9.96	11.51	12.97
	z	3.79	5.30	6.83	8.43	10.22	11.73	12.44	13.47
$\mathcal{M}(6,r_6)$	y	2.94	4.77	6.60	8.42	10.23	12.09	13.96	15.70
	z	4.25	6.07	7.88	9.66	11.52	13.04	14.62	16.29
$\mathcal{M}(7,r_7)$	y	3.77	5.86	7.97	10.08	12.12	14.63	16.51	18.55
	z	5.08	7.16	9.25	11.33	13.34	15.03	17.12	19.14
$\mathcal{M}(8,r_8)$	y	4.29	6.68	9.11	11.46	14.42	16.75	18.89	21.23
	z	5.60	7.99	10.42	12.73	14.43	17.79	20.33	24.43

Table 6: Number of correct digits (Δ) for Problem (134) with $\epsilon = 10^{-2}$ on $[0,4]$.

Method	Average order		Asymptotic order		Predicted order	
	y	z	y	z	y	z
$\mathcal{M}(2,r_2)$	2.1	1.3	2.0	1.0	2	1
$\mathcal{M}(3,r_3)$	3.0	2.4	3.0	2.1	3	2
$\mathcal{M}(4,r_4)$	4.0	3.6	4.0	2.8	4	3
$\mathcal{M}(5,r_5)$	5.0	4.6	4.9	3.4	5	4
$\mathcal{M}(6,r_6)$	6.1	5.7	5.8	5.5	6	5
$\mathcal{M}(7,r_7)$	7.0	6.7	6.8	6.7	7	6
$\mathcal{M}(8,r_8)$	8.0	8.9	7.8	13.6	8	7

Table 7: Observed orders for Problem (134).

Method / h		0.1	0.05	0.025	0.0125	0.00625	0.003125	1.562510^{-3}	7.812510^{-4}
$\mathcal{M}(2,r_2)$	y	2.67	2.93	3.36	3.96	4.65	5.35	6.02	6.66
	z	1.00	1.30	1.63	1.99	2.34	2.65	2.96	3.27
$\mathcal{M}(3,r_3)$	y	3.49	4.55	5.17	6.08	7.07	8.05	9.00	9.94
	z	2.14	2.83	3.67	4.41	5.05	5.65	6.25	6.85
$\mathcal{M}(4,r_4)$	y	3.97	5.73	6.27	7.41	8.71	10.06	11.38	12.61
	z	2.60	4.66	5.58	6.84	8.06	8.59	9.40	10.27
$\mathcal{M}(5,r_5)$	y	2.19	5.28	7.77	9.42	11.08	12.73	14.37	15.96
	z	1.06	3.88	6.95	8.85	10.48	11.31	12.43	13.60
$\mathcal{M}(6,r_6)$	y	4.16	5.55	9.27	11.00	12.85	14.85	16.82	18.75
	z	2.33	4.10	8.11	9.87	11.67	13.53	15.22	16.81
$\mathcal{M}(7,r_7)$	y	2.70	3.05	9.19	12.96	15.24	17.53	19.81	22.05
	z	1.01	0.73	8.18	11.83	14.07	16.20	18.19	20.07
$\mathcal{M}(8,r_8)$	y	3.55	3.51	10.32	14.64	17.13	19.70	22.29	24.85
	z	1.84	1.71	9.04	13.61	16.12	18.70	22.39	23.57

Table 8: Number of correct digits (Δ) for Problem (136) with $\nu = 10$ on $[0,1]$.

Method	Average order		Asymptotic order		Predicted order	
	y	z	y	z	y	z
$\mathcal{M}(2,r_2)$	2.2	1.1	2.1	1.0	2	1
$\mathcal{M}(3,r_3)$	3.2	2.0	3.1	2.0	3	2
$\mathcal{M}(4,r_4)$	4.3	2.8	4.1	2.9	4	3
$\mathcal{M}(5,r_5)$	5.4	3.9	5.3	3.9	5	4
$\mathcal{M}(6,r_6)$	6.4	5.8	6.5	5.3	6	5
$\mathcal{M}(7,r_7)$	7.5	6.8	7.4	6.2	7	6
$\mathcal{M}(8,r_8)$	8.5	8.3	8.5	3.9	8	7

Table 9: Observed orders for Problem (136).

Unité de Recherche INRIA Rennes
IRISA, Campus Universitaire de Beaulieu 35042 RENNES Cedex (France)

Unité de Recherche INRIA Lorraine Technopôle de Nancy-Brabois - Campus Scientifique
615, rue du Jardin Botanique - B.P. 101 - 54602 VILLERS LES NANCY Cedex (France)
Unité de Recherche INRIA Rhône-Alpes 46, avenue Félix Viallet - 38031 GRENOBLE Cedex (France)
Unité de Recherche INRIA Rocquencourt Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 LE CHESNAY Cedex (France)
Unité de Recherche INRIA Sophia Antipolis 2004, route des Lucioles - B.P. 93 - 06902 SOPHIA ANTIPOLIS Cedex (France)

EDITEUR
INRIA - Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 LE CHESNAY Cedex (France)

ISSN 0249 - 6399